

1/4

PREPARATION PHASE

24



USER DEFINES THE FOLLOWING:

WEB PAGE CONTENT TYPES  
THAT THE METHOD MUST  
RECOGNIZE

SET OF TESTS THAT PROVIDE  
EVIDENCE ABOUT THE  
CONTENT TYPE

N (COMPANY NEWS)  
C (CONTACT INFORMATION)  
P (PRODUCT INFORMATION)  
M (MANAGEMENT TEAM)  
D (COMPANY DESCRIPTION)  
...etc...

T1 = "NUMBER OF EXTERNAL  
LINKS ON PAGE > 5"  
T2 = "NUMBER OF INTERNAL  
LINKS > 10"  
T3 = "LINK TEXT CONTAINS  
CONTACT KEYWORDS  
(e.g. ADDRESS, LOCATION,  
CONTACT, etc)"  
T4 = "NUMBER OF PEOPLE  
NAMES IN PAGE > 3"  
T5 = "PAGE CONTAINS  
STOCK TICKER SYMBOL"  
T6 = "PAGE CONTAINS  
HEADER STARTING  
WITH WORD "ABOUT.."  
...etc..."

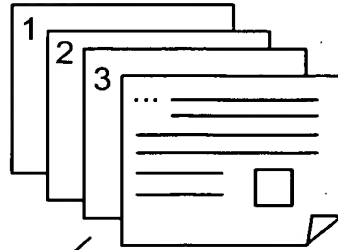
FIG. 1

2/4

TRAINING PHASE

50

23  
 TRAINING SET OF WEB  
 PAGES WITH KNOWN  
 CONTENTS



20  
 CONTENT TYPES FOR  
 EACH WEB PAGE IN  
 THE TRAINING SET

PAGE	CONTENT TYPES
1	N, C, P
2	N, C
3	D, M
4	M, P, C
.....etc....	

22  
 TEST RESULTS FOR EACH  
 WEB PAGE IN THE  
 TRAINING SET

PAGE	T1	T2	T3	T4
1	T	F	T	F
2	F	T	F	F
3	F	F	T	T
4	F	F	T	T

CALCULATE  
 STATISTICS

$P(H=N) = 0.20$	$P(H=C) = 0.20$	.....etc.....
$P(T1=T/H=N) = 0.4630$	$P(T1=T/H=C) = 0.2344$	
$P(T1=F/H=N) = 0.5370$	$P(T1=F/H=C) = 0.7656$	
$P(T2=T/H=N) = 0.2647$	$P(T2=T/H=C) = 0.6224$	.....etc.....
$P(T2=F/H=N) = 0.7353$	$P(T2=F/H=C) = 0.3776$	
$P(T3=T/H=N) = 0.7352$	$P(T3=T/H=C) = 0.2432$	
$P(T3=F/H=N) = 0.2648$	$P(T3=F/H=C) = 0.7568$	
.....etc.....	.....etc.....	

FIG. 2

CLASSIFICATION PHASE

52

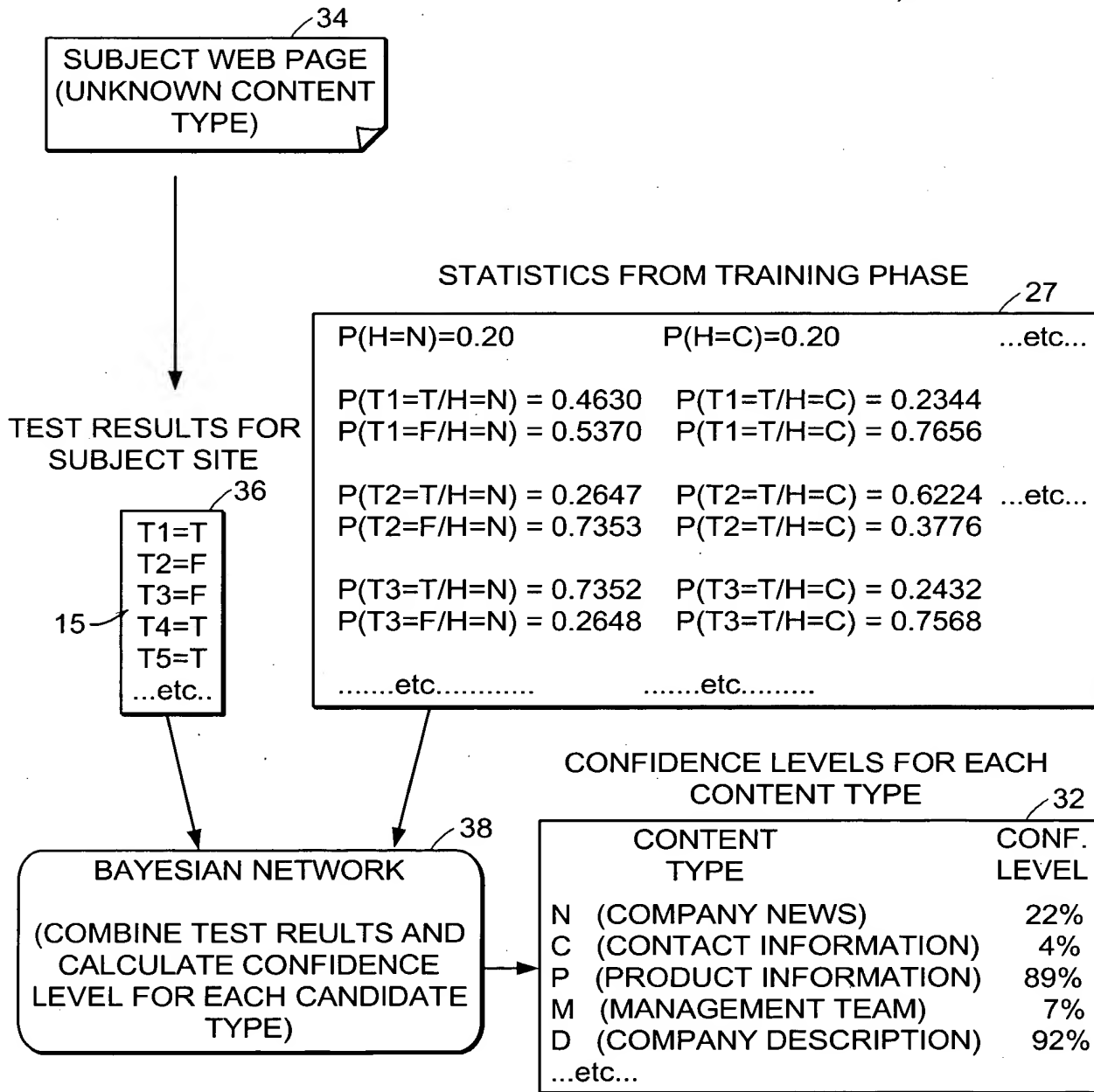


FIG. 3

PREFERRED EMBODIMENT

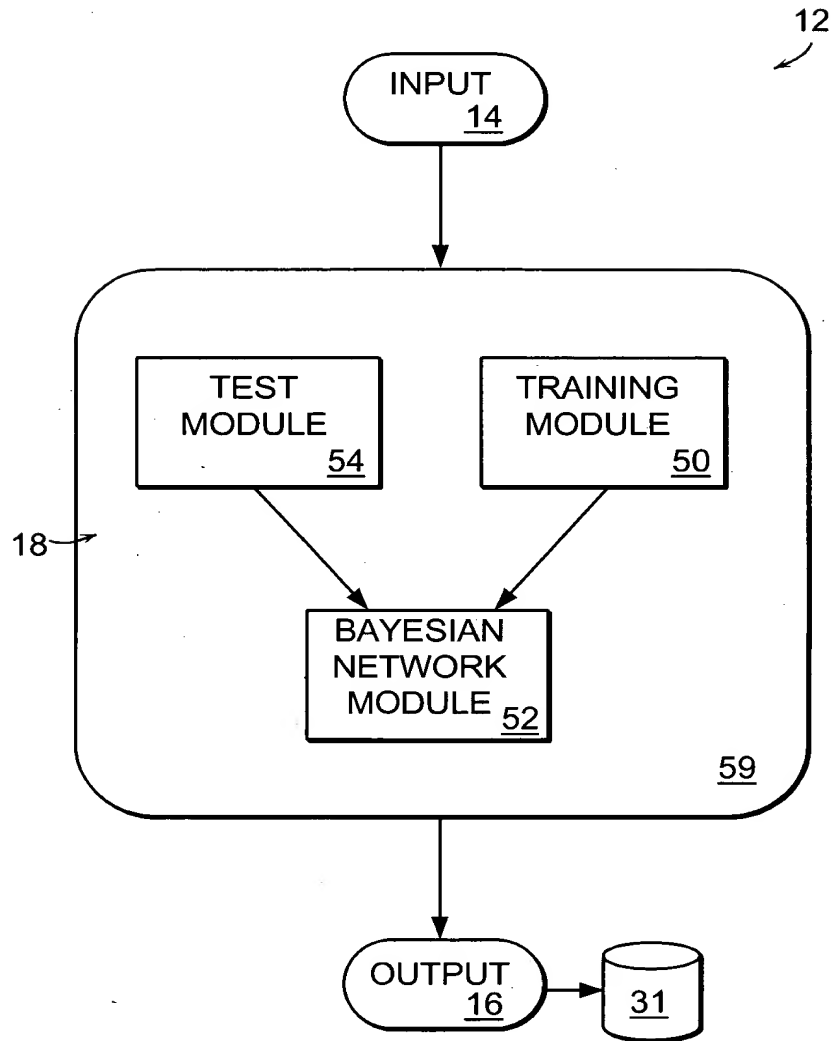


FIG. 4

09768868-041601